

Reference 1

PATENT ABSTRACTS OF JAPAN

(11)Publication number : 06-348755

(43)Date of publication of application : 22.12.1994

(51)Int.Cl.

G06F 15/40

(21)Application number : 05-135588

(71)Applicant : HITACHI LTD

(22)Date of filing : 07.06.1993

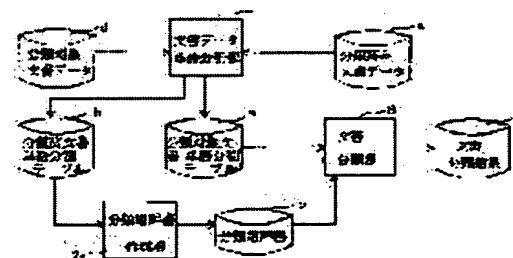
(72)Inventor : KIYAMA TADAHIRO
TSUJI HIROSHI

(54) METHOD AND SYSTEM FOR CLASSIFYING DOCUMENT

(57)Abstract:

PURPOSE: To save huge human labor for classifying document data by preparing any specified dictionary for classification and classifying non-classified documents while utilizing this dictionary for classification.

CONSTITUTION: A document data word division part 1 divides the document data into words while referring to classified document data (a) and registers the classified result on a classified word division table (b). A dictionary for classification preparation part 2 detects the appearance frequency of words while referring to the classified word division table (b) and registers the result on a dictionary (c) for classification. After the dictionary (c) is generated, the word division part 1 divides the document data into words while referring to classification target document data (d) and registers the divided result on a classification target word division table (e). A document classification part 3 classifies the document data (e) while referring to the division table (e) and the dictionary (c) for classification and registers them on a document classified result (f). Thus, since the system is provided with the dictionary for classification preparing function and the document classifying function, the document data can be automatically classified.



LEGAL STATUS

[Date of request for examination] 18.05.2000

[Date of sending the examiner's decision of rejection] 07.01.2003

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

(19) 日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11) 特許出願公開番号

特開平6-348755

(43) 公開日 平成6年(1994)12月22日

(51) Int.Cl.⁵

G 0 6 F 15/40

識別記号

5 0 0 T

庁内整理番号

9194-5L

F I

技術表示箇所

審査請求 未請求 請求項の数12 O L (全 20 頁)

(21) 出願番号 特願平5-135588

(22) 出願日 平成5年(1993)6月7日

(71) 出願人 000005108

株式会社日立製作所

東京都千代田区神田駿河台四丁目6番地

(72) 発明者 木山 忠博

神奈川県川崎市麻生区王禅寺1099番地 株式会社日立製作所システム開発研究所内

(72) 発明者 辻 洋

神奈川県川崎市麻生区王禅寺1099番地 株式会社日立製作所システム開発研究所内

(74) 代理人 弁理士 小川 勝男

(54) 【発明の名称】 文書分類方法およびそのシステム

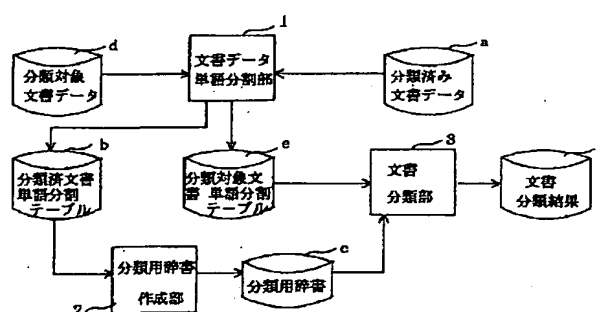
(57) 【要約】

【目的】 人手による文書データの分類作業負担を軽減するために、複数の分類の文書データとを利用し、分類別のキーワードを抽出し分類用辞書を作成し、分類用辞書を利用して文書データを自動的に分類する方法およびシステムを提供することを目的とする。

【構成】 1は、aを参照し、分類済文書データを単語分割し、bに登録する。また、1はdを参照し、分類対象文書データを単語分割し、eに登録する。2は、bを参照し、分類別のキーワードを検出し、cに登録する。3は、eとcを参照し、分類対象文書を分類し、fに登録する。

【効果】 従来は人手により分類されていた文書データを自動的に分類することが可能となり、人手による文書データの分類作業に費やす膨大な作業を省くことができるようになるという効果がある。

図1



【特許請求の範囲】

【請求項1】文書データを分類する処理システムにおいて、

一分類が一文書データ以上からなる分類済みの文書データから分類別のキーワードとなる語を抽出し分類用辞書を作成する手段と、

前記分類用辞書を用いて未分類の文書データを分類する手段を有する文書分類システム。

【請求項2】請求項1記載の文書分類システムにおいて、

分類用辞書を作成する手段は、

分類済みの文書データを用いて前記文書データ内の単語を検出し唯一の分類のみに出現する単語を検出し、前記単語を前記分類を表わすキーワードとして分類用辞書に登録することを特徴とする文書分類システム。

【請求項3】請求項1記載の文書分類システムにおいて、

分類用辞書を作成する手段は、

分類済みの文書データを用いて前記文書データを構成する項目を検出し項目別の単語を検出し唯一の分類のみに出現する単語を検出し、該単語は前記分類を表わし前記項目に出現するキーワードとして分類用辞書に登録することを特徴とする文書分類システム。

【請求項4】請求項1記載の文書分類システムにおいて、

分類用辞書を作成する手段は、

前記分類済み文書データを構成する項目が存在し利用者により前記分類用辞書を作成する範囲を前記項目により指定された場合に、前記分類済み文書データ中の指定された項目に相当する文書データのみを対象として前記分類用辞書を作成することを特徴とする文書分類システム。

【請求項5】請求項1記載の文書分類システムにおいて、

未分類の文書データを分類する手段は、

前記未分類文書データ中の単語を検出し前記分類用辞書に登録済みのキーワードとの一致数を検出し一致した分類の中で最も一致数が多い分類を前記未分類文書データの分類結果とすることを特徴とする文書分類システム。

【請求項6】請求項1記載の文書分類システムにおいて、

未分類の文書データを分類する手段は、

前記未分類文書データ中の単語を検出し前記分類用辞書に登録済みのキーワードとの一致数を検出し一致した分類の中で一致数が多い順番に優先度が高い分類として前記未分類文書データを分類することを特徴とする文書分類システム。

【請求項7】請求項1記載の文書分類システムにおいて、

未分類の文書データを分類する手段は、

前記未分類文書データを構成する項目が存在し利用者により前記未分類文書データを分類する範囲を前記項目により指定された場合に、前記未分類文書データ中の指定された項目に相当する文書データのみを処理の対象とすることを特徴とする文書分類システム。

【請求項8】文書データを分類する処理システムにおいて、

一分類が一文書データ以上からなる分類済みの文書データから分類別のキーワードとなる語を抽出し分類用辞書を作成する手段と、

前記分類用辞書を用いて未分類の文書データを分類する手段と、前記未分類文書データを分類する手段により分類された結果から新たにキーワードを検出し前記分類用辞書に登録する手段を有する文書分類システム。

【請求項9】請求項8記載の文書分類システムにおいて、

新たにキーワードを検出し前記分類用辞書に登録する手段は、

前記未分類文書データを分類する手段により分類された文書データ中の単語で前記分類用辞書中に存在しない単語であれば該単語は前記分類を表わすキーワードとして前記分類用辞書に登録することを特徴とする文書分類システム。

【請求項10】請求項8記載の文書分類システムにおいて、

新たにキーワードを検出し前記分類用辞書に登録する手段は、

前記未分類文書データを分類する手段による分類結果が正しいか否か利用者に問合せ正しいと指示された場合に、前記未分類文書データを分類する手段により分類された文書データ中の単語で前記分類用辞書中に存在しない単語であれば該単語を前記分類を表わすキーワードとして前記分類用辞書に登録することを特徴とする文書分類システム。

【請求項11】請求項8記載の文書分類システムにおいて、

新たにキーワードを検出し前記分類用辞書に登録する手段は、

前記未分類文書データを分類する手段による分類結果が正しいか否か利用者に問合せ正しいと指示された場合に、前記未分類文書データを分類する手段により分類された文書データ中の単語で前記分類用辞書中に存在しない単語であれば該単語を前記分類を表わすキーワードとして利用者に提示し、利用者が指示したキーワードのみを前記分類用辞書に登録することを特徴とする文書分類システム。

【請求項12】請求項8記載の文書分類システムにおいて、

新たにキーワードを検出し前記分類用辞書に登録する手段は、

前記未分類文書データを分類する手段による分類結果が正しいか否か利用者に問合せ正しいと指示された場合に一致したキーワードを利用者に提示し、誤ったキーワードが存在している場合に利用者の指示により分類用辞書から該キーワードを削除することを特徴とする文書分類システム。

【発明の詳細な説明】

【0001】

【産業上の利用分野】本発明は文書データの自動分類方法に係り、特に、大量の文書データを自動的に分類する場合に好適な文書分類方法およびそのシステムに関する。

【0002】

【従来の技術】従来の文書分類方法は、例えば、特開平2-158871号公報に記載されている。文書に含まれるキーワードの頻度値から各文書の概念特徴量を求め、これに応じて文書を分類している。

【0003】

【発明が解決しようとする課題】従来の文書分類方法では、文書間の概念特徴量の差に応じて文書間の距離を求め、この距離によって文書の分類を行なっているが、概念特徴量を求めるために、文書分類用のシソーラスやキーワード分類項目を予め人手により登録しこれを利用している。つまり、文書を分類するために分類用の情報を人手により定義しなければならないという問題があった。

【0004】本発明の目的は、上記問題を解決するために、複数に分類された書データと、分類内での各単語の出現頻度を利用し、分類別のキーワードを自動的に抽出し、分類用辞書を作成し、この分類用辞書を利用することにより未分類の文書を分類する方法およびシステムを提供することにある。

【0005】

【課題を解決するための手段】上記目的を達成するために、本発明の文書分類方法は、一分類が一文書データ以上からなる分類済みの文書データから分類別のキーワードとなる語を抽出し分類用辞書を作成する方法と、分類用辞書を用いて未分類の文書データを分類する方法と、未分類文書データを分類する方法により分類された結果から新たにキーワードを検出し分類用辞書に登録する方法を具備する。また、本発明の文書分類システムは、上記文書分類方法を実行するプログラムを具備した自然言語処理システムである。

【0006】

【作用】本発明の文書分類方法およびシステムは、まず、分類済みの文書データを取得する。次に、利用者により分類用辞書の作成範囲を文書データを構成する項目により指定されている場合は指定項目のみの文書データを処理対象とし、指定されてない場合は文書データ全体を処理対象とし分類用辞書を作成する。次に、分類済みの文書データから得た分類別に、唯一の分類のみに存在

する単語を検出し、この単語を該当する分類のキーワードとして分類用辞書に登録する。また、文書データが複数の項目により構成されている場合にはキーワードと項目の対応関係も分類用辞書登録する。

【0007】分類用辞書作成後、未分類の文書データを取得する。次に、利用者により文書データの分類範囲を文書データを構成する項目により指定されている場合は指定項目のみの文書データを処理対象とし、指定されてない場合は文書データ全体を処理対象とし文書を分類する。未分類文書データの単語を検出し、分類用辞書中のキーワードと比較照合し、分類別に一致回数を求める。

【0008】次に、キーワードの一致回数が最も多い分類を分類対象文書データの分類結果とする。また、一致回数が多い分類の順番に優先度を付与し分類結果とする。また、未分類文書データを分類する方法による分類結果が正しいか否か利用者に問合せ正しいと指示された場合に、未分類文書データを分類する方法により分類された文書データ中の単語で分類用辞書中に存在しない単語であれば、この単語を該分類を表わすキーワードとして分類用辞書に登録する。また、キーワードを分類用辞書に登録する場合に、利用者が指示したキーワードのみを分類用辞書に登録する。

【0009】また、未分類文書データを分類する方法による分類結果が正しいか否か利用者に問合せ正しいと指示された場合に一致したキーワードを利用者に提示し、誤ったキーワードが存在している場合に利用者の指示により分類用辞書から該キーワードを削除する。以上により、分類されてない文書データを自動的に分類することができると同時に、分類用辞書の自己増殖が可能となる。

【0010】

【実施例】以下、本発明の実施例を、図面により詳細に説明する。図1は、本発明の文書分類方法の一実施例を示す機能ブロック図である。文書データ単語分割部1は、分類済文書データaを参照し、文書データを単語分割し、分割結果を分類済単語分割テーブルbに登録する。分類用辞書作成部2は、分類済単語分割テーブルbを参照し、単語の出現頻度を検出し、分類用辞書cに登録する。分類用辞書が生成された後、文書データ単語分割部1は、分類対象文書データbを参照し、文書データを単語分割し、分割結果を分類対象単語分割テーブルeに登録する。文書分類部3は、分類対象単語分割テーブルeと分類用辞書を参照し、分類対象文書データdを分類し文書分類結果fに登録する。

【0011】このように分類用辞書作成機能及び文書分類機能を持たせることにより、文書データの自動分類を実現する。

【0012】図から明らかなように、文書データ単語分割部1、分類用辞書作成部2、文書分類部3は処理を示し、分類済文書データa、分類済単語分割テーブルb、

分類用辞書c、分類対象文書データd、分類対象文書単語分割テーブルe、文書分類結果fはファイル（テーブルとも呼ぶ）である。このように、本実施例によれば、各機能ブロックが、プログラム論理により構成されている。そのため、各機能ブロック単位にLSI化が可能であり、文書分類装置として、処理の高速化を図ることができる。

【0013】図2は、図1における文書分類装置の全体的なハードウェア構成を示すブロック図である。入出力装置4は、データの入力、および、各種情報の表示を行なう。プロセッサ5は、プログラムに基づき、図1における処理を実行する。記憶装置6は、図1における各種文書データaや各種プログラム等を格納する。さらに、記憶装置6は、プロセッサ5の各処理実行用のメモリであるワーキングエリアae、文書データ単語分割部格納エリア10、分類用辞書作成部格納エリア20、文書分類部格納エリア30、文書データ格納エリアad、単語分割テーブル格納エリアbe、分類用辞書格納エリアc、文書分類結果格納エリアfの記憶部を持っている。

【0014】記憶装置6に格納される各プログラムは、プロセッサ5において実行される。その実行に際して、必要に応じて入出力装置4が用いられる。

【0015】図3は、図1における文書データ単語分割部の処理手順を表すPAD図（Problem Analysis Diagram）である。分類済文書データaから文書データを取得し単語分割を行ない分類済単語分割テーブルbに格納するまでの処理、または、分類対象文書データdから文書データを取得し単語分割を行ない分類対象単語分割テーブルeに格納するまでの処理を示したものである。

【0016】以下、この処理をPAD図に従って説明する。分類済文書データaを参照し、先頭文書データから末尾文書データまで以下の処理を行なう（ステップ11）。まず、分類済文書データaから一文書分のデータを取得し（ステップ12）、文書データを単語分割し見出し文字列、品詞をそれぞれ単語分割テーブルbの見出し文字列b1、品詞b2に格納する（ステップ13）。

【0017】次に、文書データが項目で区分されているか判別し（ステップ14）、項目で区分されている場合には、各単語に該当する項目を分類済単語分割テーブルbの項目b4に格納する（ステップ15）。次に、該当する文書データの分野名、文書番号を分類済単語分割テーブルbの分類名b5、文書No b6に格納する（ステップ16）。次に、処理の対象を次の文書データに移動する（ステップ17）。

【0018】また、ステップ11～ステップ17において、分類対象文書データdを入力とした場合には、分類対象単語分割テーブルeに単語分割結果を格納する。ステップ11～ステップ17により、図4に示す分類済文書データaを取得し単語分割を行ない、図5に示す分類済単語分割テーブルbに格納する。また、ステップ11

～ステップ17により、図9に示す分類対象文書データdを取得し単語分割を行ない、図10に示す分類対象単語分割テーブルeに格納する。

【0019】図4は、分類済文書データaの例であり、分類名、文書番号付きの分類済みの文書データaの例である。分類済み文書データは「題名」「要旨」「目的」「主内容」「今後の課題」などの項目により構成されている。

【0020】図5は、単語分割テーブルの例であり、見出し文字列b1、品詞b2、項目b3、分類名b4、文書No b5の項目により構成されている。見出し文字列b1は分割された単語を構成する文字列、品詞b2は単語の品詞、項目b3は該当する単語に対応する項目の項目名、分類名b4は該当する文書データの分類名、文書No b5は文書データの識別番号を表している。

【0021】図6は、図1における分類用辞書作成部2の処理を表わすPAD図であり、未分類の文書を分類するための辞書を作成する処理を示したものである。以下、この処理をPAD図に従って説明する。

【0022】分類済単語分割テーブルbを参照し、分類済単語分割テーブルの先頭レコードから末尾レコードまで以下の処理を行なう（ステップ201）。まず、分類済単語分割テーブルbを参照し、1レコード分の情報を取得し（ステップ202）、見出し文字列b1と品詞b2と分類名b4が等しい他のレコードを分類済み単語分割テーブルbより検索し保持しワークテーブルに格納する（ステップ203）。

【0023】次に、該当する文書データを構成する項目が存在するか判別し（ステップ204）、項目が存在する場合に、項目別の各単語の出現回数を求めワークテーブルに格納する（ステップ205）。次に、文書データ全体の各単語の出現回数を求めワークテーブルに格納する（ステップ206）。

【0024】次に、処理対象となるレコードを次のレコードへ移動する（ステップ207）。ステップ201の繰返し処理が終了した後、201～206で格納したワークテーブルの先頭レコードから末尾レコードまで以下の処理を行なう（ステップ208）。

【0025】まず、ワークテーブルより1レコード文の情報を取得し保持する（ステップ209）。次に、見出し文字列と品詞が等しい他のレコードがワークテーブル中に存在するか判別し（ステップ210）、存在しない場合に、保持した本レコードを分類用辞書3に格納する（ステップ211）。

【0026】次に、処理対象となるレコードを次のレコードへ移動する（ステップ212）。

【0027】ステップ201～ステップ207により、図5に示す分類済単語分割テーブルbから図7に示すワークテーブルが生成され、ステップ208～ステップ212により、図7に示すワークテーブルから図8に示す

分類用辞書cが生成される。

【0028】図7は、図6におけるワークテーブルの例であり、見出し文字列W1、品詞W2、題名W3～今後の課題W7、合計W8、分類名W9の項目により構成されている。見出し文字列W1は分割された単語を構成する文字列、品詞W2は単語の品詞、題名W3～今後の課題W7は文書データを構成する項目中に出現する単語の出現回数、合計W8は文書データ中に出現する単語の出現回数、分類名W9は該当する文書データの分類名を表している。

【0029】図8は、図1における分類用辞書の例であり、見出し文字列C1、品詞C2、題名C3～今後の課題C7、合計C8、分類名C9の項目により構成されている。見出し文字列C1は分割された単語を構成する文字列、品詞C2は単語の品詞、題名C3～今後の課題C7は文書データを構成する項目中に出現する単語の出現回数、合計C8は文書データ中に出現する単語の出現回数、分類名C9は該当する文書データの分類名を表している。本分類用辞書は図7に示すワークテーブルとは異なり、見出し文字列が重複することなく、一つの見出し文字列は一つの分類名に対応している。

【0030】図9は、分類対象文書データaの例であり、文書番号付きの分類対象文書データdの例である。分類対象文書データは「題名」「要旨」「目的」「主内容」「今後の課題」などの項目により構成されている。

【0031】図10は、分類対象文書単語分割テーブルdの例であり、見出し文字列e1、品詞e2、項目e3、分類名e4、文書Noe5の項目により構成されている。見出し文字列e1は分割された単語を構成する文字列、品詞e2は単語の品詞、項目e3は該当する項目の項目名、分類名e4は該当する文書データの分類名、文書Noe5は文書データの識別番号を表している。

【0032】図11は、図1における文書分類部3の処理を表すPAD図である。本処理は、未分類の文書を分類用辞書を用いて分類する処理である。また、本処理の前提として、図1における分類対象文書データdを文書データ単語分割部1により単語分割し、単語分割結果を分類対象文書単語分割テーブルeに格納されているものとする。

【0033】以下、この処理をPAD図に従って説明する。分類対象文書単語分割テーブルbを参照し、本テーブルの先頭レコードから末尾レコードまで以下の処理を行なう（ステップ301）。まず、利用者が文書の分類の対象とする項目を指定しているか判別し（ステップ302）、指定していれば、該当する項目に対応するレコードに限定しレコード分の情報を取得し（ステップ303）、指定しなければ、一レコード分の情報を取得する（ステップ304）。

【0034】次に、取得したレコードを文字列変数MIDASHI1に格納する（ステップ305）。次に、分

類用辞書cを参照し、分類用辞書の先頭レコードから末尾レコードまで以下の処理を行なう（ステップ306）。まず、利用者が文書の分類の対象とする項目を指定しているか判別し（ステップ307）、指定していれば、該当する項目に対応するレコードに限定し分類用辞書cから一レコード分の情報を取得し（ステップ308）、指定しなければ、一レコード分の情報を取得する（ステップ309）。

【0035】次に、取得したレコードを文字列変数MIDASHI2に格納する（ステップ310）。次に、MIDASHI1とMIDASHI2に格納した情報から見出し文字列と品詞を取得し、見出し文字列と品詞が一致するか判別し（ステップ311）、一致する場合、一致回数を分類名別にカウントしこれを保持し（ステップ312）、一致した見出し文字列と品詞を保持する（ステップ313）。

【0036】次に、本繰返し処理の対象を分類用辞書の次のレコードへ移動する（ステップ314）。次に、ステップ306の繰返し処理が終了した後に、ステップ301の繰返し処理の対象を分類対象文書単語分割テーブルの次のレコードへ移動する（ステップ315）。次に、ステップ301の繰返し処理が終了した後に、上記で保持した一致回数が最も多い分類名を本文書の分類結果として文書分類結果fに格納し（ステップ316）、一致回数が1回以上の分類名を対象に分類名別に一致回数が多い順番に並べ換え、分類候補として文書分類結果fに格納し（ステップ317）、一致した見出し文字列と品詞と一致回数を文書分類結果fに格納する（ステップ318）。

【0037】ステップ301～ステップ318により、図10に示す分類対象文書単語分割テーブルeと分類用辞書cをもとに文書を分類し図12に示す文書分類結果fを得ることができる。この文書分類結果fを利用者に示すことにより分類結果の確認が可能となる。

【0038】図12は、図1における文書分類結果fの例である。本例には、3つの文書の分類結果を示している。1文書目は図9に示している文書No「101」の文書であり、結果として「言語処理」に分類されている。これは分類名「言語処理」が分類用辞書の単語の一致回数が最も高かったことを表わしている。また、一致した単語は「辞書」「自然語」「LR解析」等であり括弧内の数値は単語の一致回数を表わしている。また、「言語処理」の次に単語の一致回数が多かった分類候補が「通信」であることを表わしている。

【0039】2文書目は図9に示している文書No「102」の文書であり、結果として「電気回路」に分類されている。これは分類名「電気回路」が分類用辞書の単語の一致回数が最も高かったことを表わしている。また、一致した単語は「AD変換」「アナログ」「ディジタル」等であり括弧内の数値は単語の一致回数を表わし

ている。また、「電気回路」の次に単語の一致回数が多かった分類候補は無かったことを表わしている。

【0040】3文書目は図9に示している文書No「103」の文書であり、結果として「通信」に分類されている。これは分類名「通信」が分類用辞書の単語の一致回数が最も高かったことを表わしている。また、一致した単語は「ネットワーク」「プロトコル」「LAN」等であり括弧内の数値は単語の一致回数を表わしている。また、「通信」の次に単語の一致回数が多かった分類候補が「電気回路」であることを表わしている。

【0041】以上、述べたように、本実施例によれば、従来は人手により分類されていた文書データを自動的に分類することが可能となり、人手による文書データの分類作業に費やす膨大な作業を省くことができるようになるという効果がある。また、パソコンやユニックスなどのニュースサービス（掲示板）は文書の内容によって投稿すべきニュースグループが非常に多く、どのニュースグループに投稿すべきであるか判断できないことがあるが、本実施例によれば、ニュースグループにより分類されている文書から分類用辞書を作成し自動分類可能となり、投稿対象文書のニュースグループを指定しなくとも自動的に投稿すべきニュースグループを知ることができる。また、自動的に分類し投稿することができる。

【0042】次に、本発明の文書分類方法の拡張例について説明する。図13は、本発明の文書分類方法において分類した結果を利用者に確認し、確認した結果から新たにキーワードとして分類用辞書に登録することを実現する機能ブロック図である。文書分類確認部4は文書分類結果fを参照し、分類した文書の分類結果や一致したキーワードを利用者に提示し、利用者の確認を促し、一致したキーワードで相応しく無いものがあれば、そのキーワードを利用者の指定により分類用辞書から削除し、分類した文書に新たなキーワードが存在するか判別し、存在する場合には新たなキーワードを分類用辞書に登録し、分類用辞書の自己増殖を行なう。

【0043】図14の【画面1】は、分類対象文書の分類結果の確認例である。提示内容として、文書番号、題名、分類結果、分類候補がある。また、分類結果と同時に、分類用辞書中のキーワードと一致したキーワードとキーワードの一致回数を示している。これらの情報から利用者が分類結果が正しいか否か判断し、分類結果の正否を入力する。

【0044】図14の【画面2】は、【画面1】において分類結果が正しいと利用者が判断した後、分類の根拠となった分類用辞書中の該分類のキーワードと一致したキーワードを利用者に提示し、キーワードとして相応しく無いものがあれば、利用者に指定するよう促している例である。また、指定された場合には、そのキーワードを分類用辞書から削除し、分類用辞書の精度を向上させる。本例では、そのようなキーワードは無いと利用者が

応答している例を表わしている。

【0045】図15の【画面3】は、分類結果が正しい場合に、分類用辞書には存在せず、分類対象文書にのみにキーワードが存在する場合に、このキーワードを該分類の新たなキーワードとして自動登録するか否か、また、キーワードを確認した後に、利用者がキーワードを種々選択し登録するか否か利用者の応答を促している例である。本例では、新たに出現したキーワードは確認せず自動登録するよう利用者が応答している例を表わしている。

【0046】以上、述べたように、本拡張例によれば、分類用辞書中の任意の分類として誤ったキーワードが登録されている場合に、分類対象文書と一致したキーワードを利用者が確認し、誤ったキーワードを削除することにより、分類用辞書の精度向上を容易に実現することができる。また、分類対象文書に新たなキーワードが出現している場合に、そのキーワードを分類用辞書に自動登録する、または、利用者に提示し必要なキーワードのみ分類用辞書に登録することにより、利用者が時間をかけて辞書登録することなく、簡単に分類用辞書へのキーワードの辞書登録または自己増殖が可能となる。

【0047】次に、本発明の文書分類方法の別の拡張例について説明する。図16は、分類用辞書を作成する場合、または、分類対象文書を分類する場合に、文書データが一定の項目により区分されているならば、文書データ全体を分類用辞書作成または分類対象にするのではなく、文書内の項目を利用者が指定することを可能とし、その指定された項目のみを処理の対象とすることを実現する機能ブロック図である。文書分類対象項目指定部6は利用者が処理対象として指定した項目を認識し、分類用辞書を作成する場合には分類済文書データaを参照し文書データから該当する項目の文書データを取得し、文書データを分類する場合には分類対象文書データdを参照し文書データから該当する項目の文書データを取得し、以降の処理で取得した文書データのみを処理対象とする。

【0048】また、単に項目を指定するのではなく、任意の項目間の論理関係の指定及び認識を可能とすることにより、例えば、項目「主内容」または項目「今後の課題」に出現する単語、項目「主内容」と項目「今後の課題」の両項目に出現する単語を処理対象とすることが容易に実現可能である。

【0049】以上、述べたように、本拡張例によれば、文書データ全体を処理対象とせず項目を限定することから、文書データの性格によっては分類用辞書精度向上や分類精度向上が期待できると同時に、膨大な量の文書データや文書データ自体が大きい場合にも実用的な処理速度を維持することが可能となり、処理速度が向上するという効果がある。

【0050】

【発明の効果】本発明によれば、従来は人手により分類されていた文書データを自動的に分類することが可能となり、人手による文書データの分類作業に費やす膨大な作業を省くことができるようになるという効果がある。また、パソコンやユニックスなどのニュースサービス（掲示板）は文書の内容によって投稿すべきニュースグループが非常に多く、どのニュースグループに投稿すべきであるか判断できないことがあるが、本実施例によれば、ニュースグループにより分類されている文書から分類用辞書を作成し自動分類可能となり、投稿対象文書のニュースグループを指定しなくとも自動的に投稿すべきニュースグループを知ることができる。また、自動的に分類し投稿することができる。また、分類用辞書へのキーワード登録が簡単に行なうことができると同時に、分類用辞書の自己増殖も可能となるという効果がある。

【図面の簡単な説明】

【図1】本発明を施した文書分類方法の一実施例を示す機能ブロック図。

【図2】図1における文書分類方法の全体的なハードウェア構成を示すブロック図。

【図3】図1における文書データ単語分割プログラムのPAD図。

【図4】図1における分類済文書データの例。

【図5】図1における分類済単語分割テーブルの例。

【図6】図1における分類用辞書作成プログラムのPAD図。

【図7】図1における分類用辞書を作成するためのワークテーブルの例。

【図8】図1における分類用辞書の例。

【図9】図1における分類対象文書データの例。

【図10】図1における分類対象文書単語分割テーブルの例。

【図11】図1における文書分類プログラムのPAD図。

【図12】図1における文書分類結果例。

【図13】文書分類結果の確認方法及び分類用辞書学習方法を示す機能ブロック図。

【図14】図13における分類結果確認及び分類用辞書学習における利用者との対話例。

【図15】図13における分類結果確認及び分類用辞書学習における利用者との対話例。

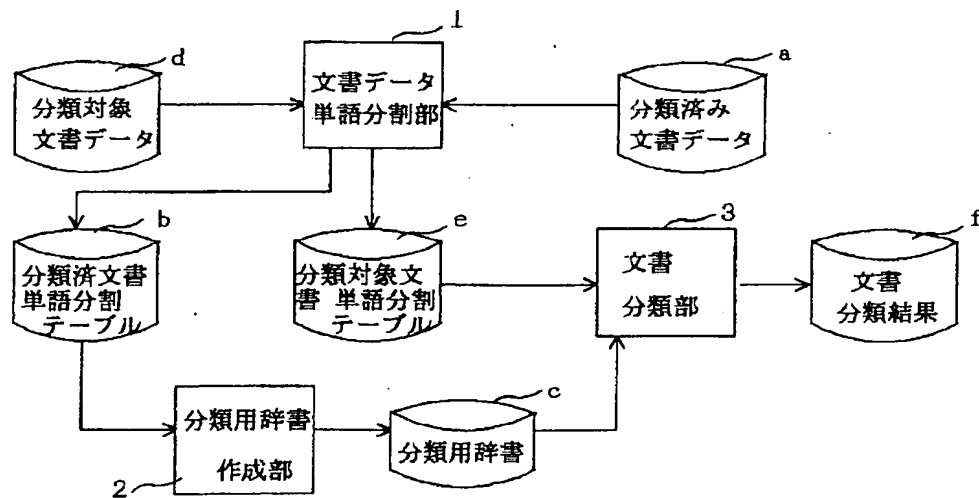
【図16】文書分類方法において処理対象とする文書内の項目指定方法を示す機能ブロック図である。

【符号の説明】

1…文書データ単語分割部、2…分類用辞書作成部、3…文書分類部、a…分類済文書データ、b…分類済単語分割テーブル、c…分類用辞書、d…分類対象文書データ、e…分類対象文書単語分割テーブル、f…分類対象文書単語分割テーブル。

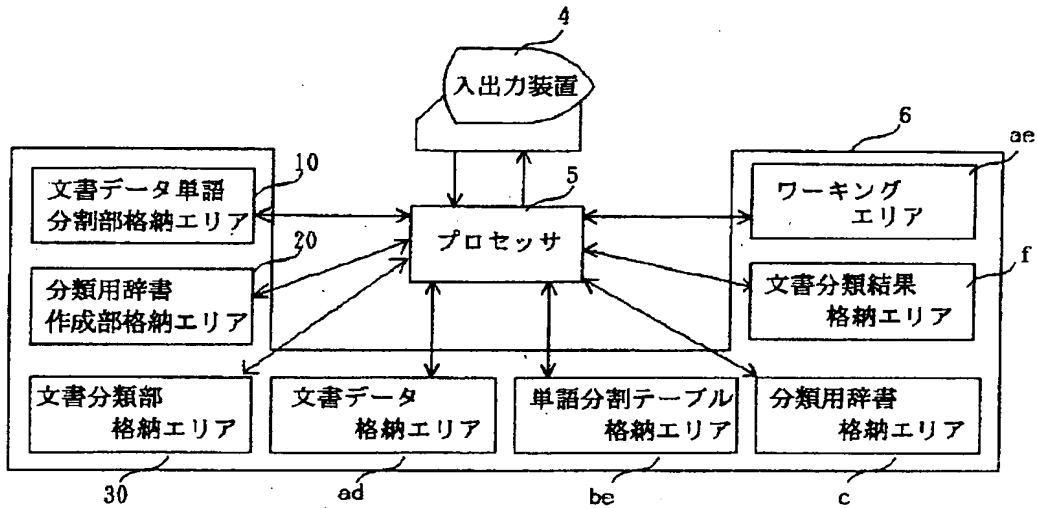
【図1】

図1



【図2】

図2



【図3】

図3

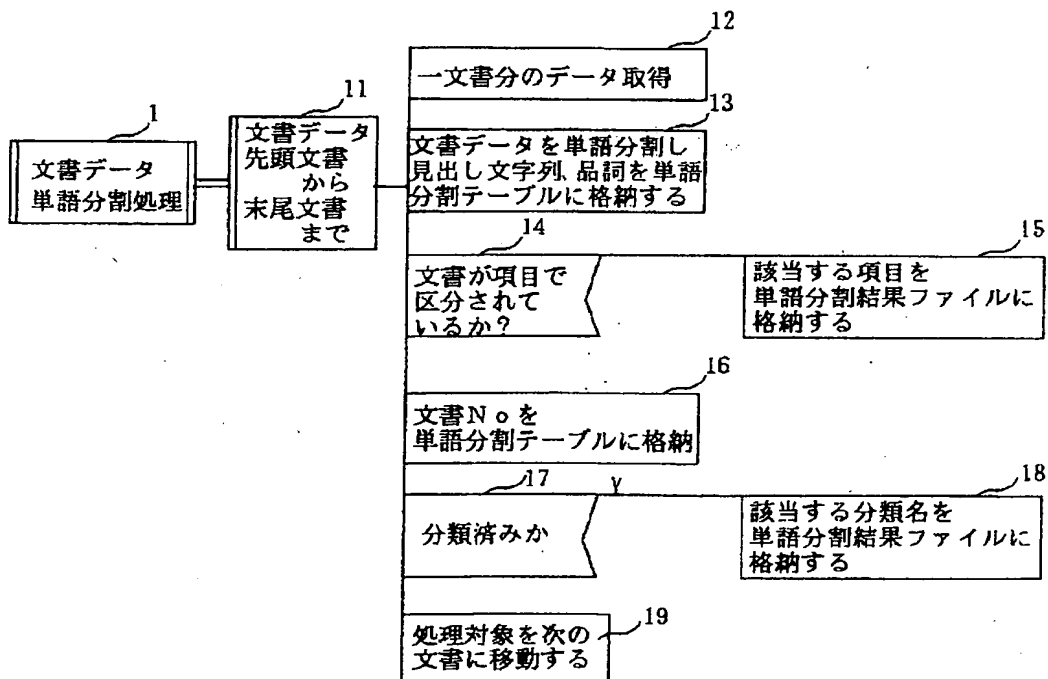


图 4

{分類名}: 言語処理
{分題名}: 1

<主題名> 生成方法に関する一考察

<要旨> 自然言語処理の発展に伴い、日本語の処理技術が飛躍的に進歩している。本論文は、日本語の形態素解析、句構解析、意味解析の各段階における最新の研究成果をレビューし、今後の研究課題を提示する。

<目的> 日本語の自然言語処理技術の現状と将来の展望を明らかにする。

<主内容> 1. 形態素解析の技術と応用
2. 句構解析の技術と応用
3. 意味解析の技術と応用

<今後の課題> 1. 大規模な日本語コーパスの構築
2. 自然言語処理の応用分野の拡大

{分類名}: 電気回路
{分題名}: 2

<主題名> 変換の拡張

<要旨> アログからディジタルへの変換技術の進展について述べる。

<目的> アログ回路からディジタル回路への変換技術の進展を調査し、その応用範囲を拡大すること。

<主内容> 1. アログ回路からディジタル回路への変換技術の進展
2. ディジタル回路の応用範囲の拡大

<今後の課題> 1. ディジタル回路の応用範囲の拡大
2. ディジタル回路の応用範囲の拡大

{分類名}: 通信
{分題名}: 3

<主題名> コンLANの実例

<要旨> 高度情報通信時代を向かへる。

<目的> 高度情報通信時代の到来に伴い、コンピュータネットワークの応用範囲が拡大している。本論文は、コンピュータネットワークの応用範囲を拡大すること。

<主内容> 1. コンピュータネットワークの応用範囲の拡大
2. コンピュータネットワークの応用範囲の拡大

<今後の課題> 1. コンピュータネットワークの応用範囲の拡大
2. コンピュータネットワークの応用範囲の拡大

【図5】

図5

b 1	b 2	b 3	b 4	b 5
見出し文字列	品 詞	項 目	分類名	文書No
題名	名詞	題名	言語処理	1
辞書	名詞	題名	言語処理	1
生成	名詞	題名	言語処理	1
方法	名詞	題名	言語処理	1
関	動詞	題名	言語処理	1
一	名詞	題名	言語処理	1
考察	サ変動詞	要旨	言語処理	1
要旨	名詞	要旨	言語処理	1
自然語	名詞	要旨	言語処理	1
インタフェース	名詞	要旨	言語処理	1
参照	サ変動詞	要旨	言語処理	1
.
.
AD変換	名詞	題名	電気回路	2
拡張	サ変名詞	題名	電気回路	2
アナログ	名詞	要旨	電気回路	2
ディジタル	名詞	要旨	電気回路	2
周波数	名詞	主内容	電気回路	2
.
.
パソコン	名詞	題名	通信	3
LAN	動詞	題名	通信	3
実例	名詞	題名	通信	3
.
.
.

【図6】

図6

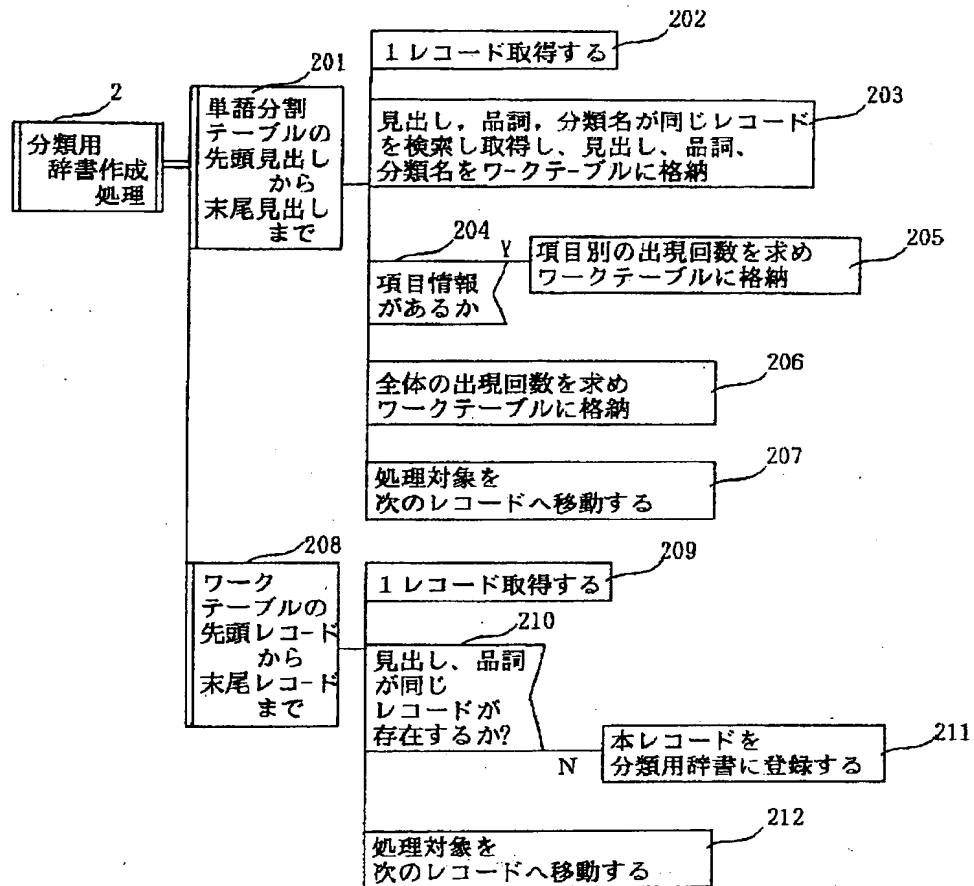


图 7

[illegible]

圖 8

[illegible]

【図9】

図9

{文書No} : 101
 <題名> 翻訳用辞書構ツールの開発
 <要旨> 機械翻訳用辞書の構築
 <目的> 自然語処理技術の研究において、機械翻訳用の辞書を構築するための
 <主内容> 辞書の開発を行った。
 <今後の課題>

{文書No} : 102
 <題名> 変換の拡張 (その2)
 <要旨> A/D変換
 <目的> 前回のアナログからデジタルへ
 <主内容>
 <今後の課題>

{文書No} : 103
 <題名> ネットワークを用いたデータベース検索
 <要旨> ネットワークを用いたデータベース検索システムを開発した。本報では
 <目的> ネットワークの
 <主内容>
 <今後の課題>

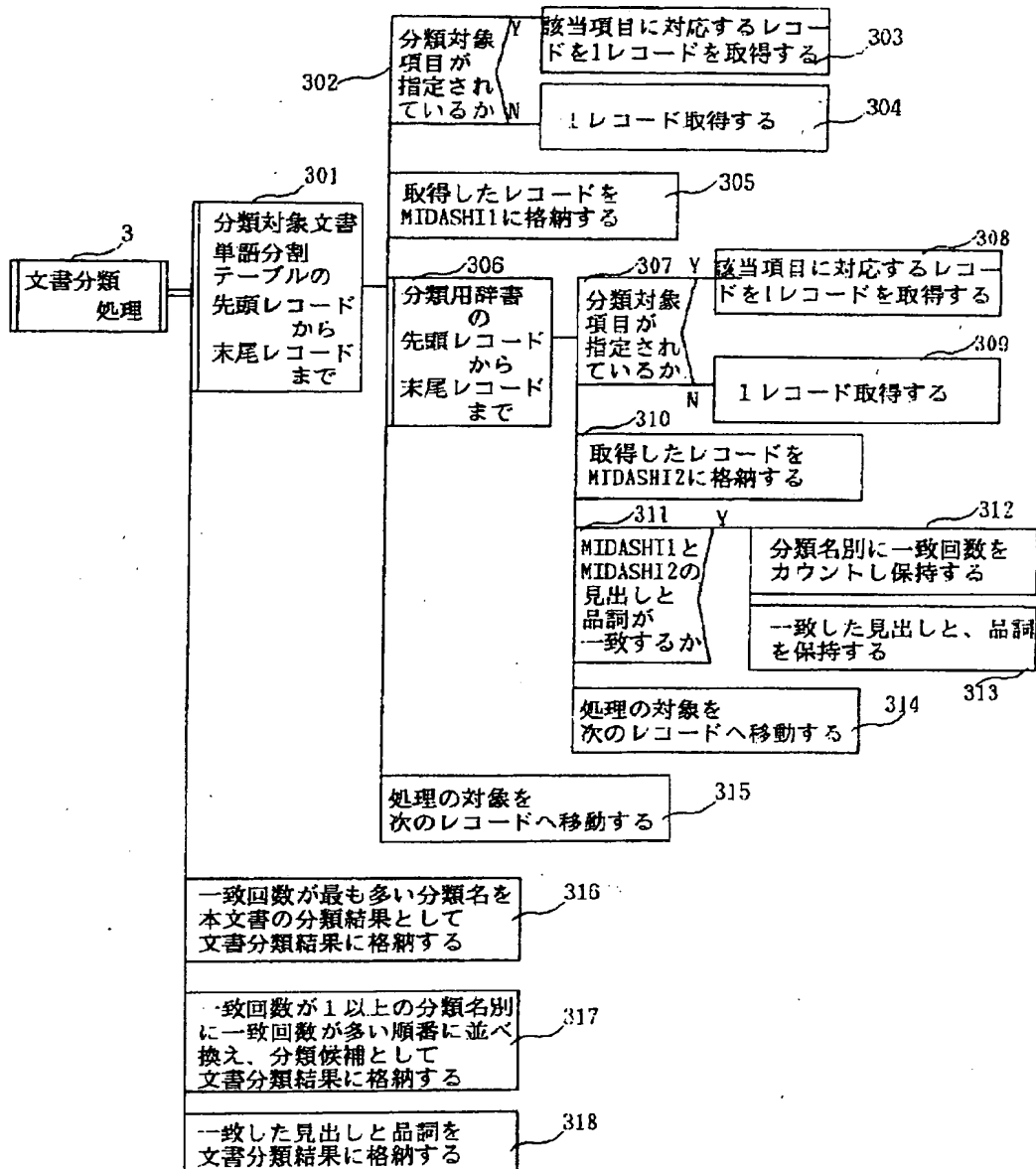
【図10】

図10

e 1 見出し文字列	e 2 品 詞	e 3 項 目	e 4 分類名	e 5 文書No
題名	名詞	題名	なし	101
機械翻訳	名詞	題名	なし	101
用	名詞	題名	なし	101
辞書	名詞	題名	なし	101
構築	サ変動詞	題名	なし	101
ツール	名詞	題名	なし	101
開発	サ変動詞	要旨	なし	101
要旨	名詞	要旨	なし	101
自然語	名詞	要旨	なし	101
処理	サ変動詞	要旨	なし	101
技術	名詞	要旨	なし	101
LR解析	名詞	要旨	なし	101
.
.
AD変換	名詞	題名	なし	102
拡張	サ変名詞	題名	なし	102
その	代名詞	題名	なし	102
2	数字	題名	なし	102
前回	名詞	要旨	なし	102
アナログ	名詞	要旨	なし	102
デジタル	名詞	要旨	なし	102
.
.
ネットワーク	名詞	題名	なし	103
用い	動詞	題名	なし	103
データベース	名詞	題名	なし	103
検索	名詞	題名	なし	103
プロトコル	名詞	要旨	なし	103
周波数	名詞	主内容	なし	103
.
.
.

【図11】

図11



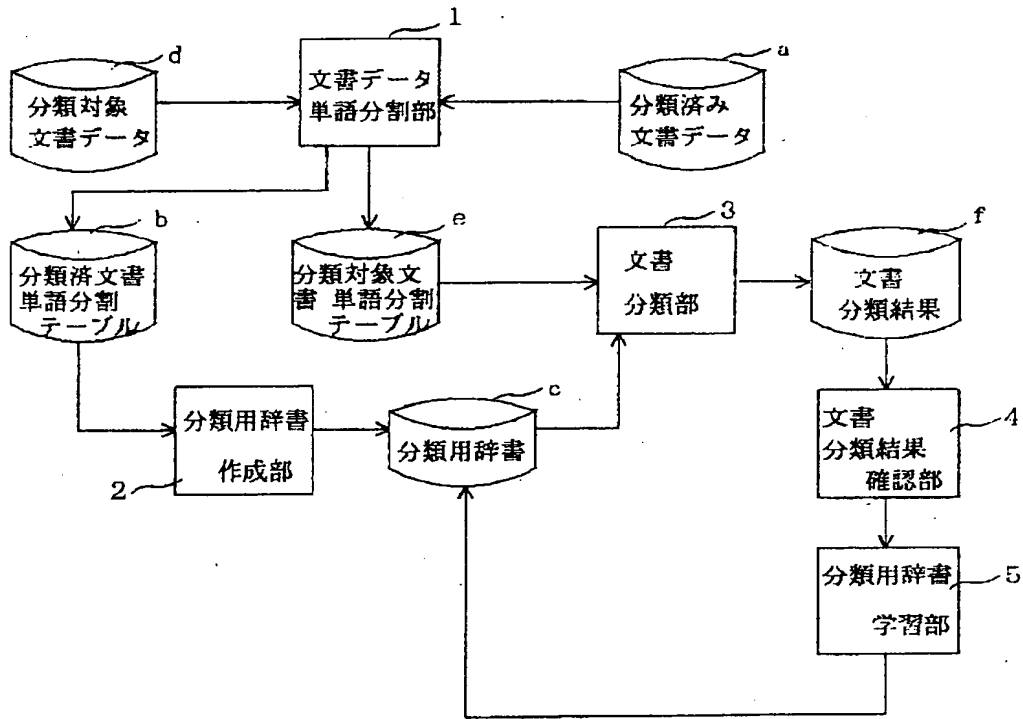
【図12】

図12

【文書No.】	: 101
【題名】	: 機械翻訳用辞書構築ツールの開発
【分類結果】	: 言語処理(辞書(25),自然語(23),LR解析(6)...)
【分類候補】	: 通信(ネットワーク(3))
【文書No.】	: 102
【題名】	: AD変換の拡張(その2)
【分類結果】	: 電気回路(AD変換(35),アナログ(31),ディジタル(30)...)
【分類候補】	: なし
【文書No.】	: 103
【題名】	: ネットワークを用いたデータベース検索
【分類結果】	: 通信(ネットワーク(40),プロトコル(25),LAN(11)...)
【分類候補】	: 電気回路(周波数(2))

【図13】

図13



【図15】

図15

【画面3】

<div style="display: flex; justify-content: space-around; margin-bottom: 10px;"> 登録しない 確認後登録 自動登録 </div> <div style="text-align: center;"> </div>	
<p>本分類結果をもとに、分類用辞書に新たにキーワードを登録できます。 キーワードの登録方法として3通りあります。 該当する登録方法をマウスで指定して下さい。</p>	
[文書No.]	: 1'01
[題名]	: 機械翻訳用辞書構築ツールの開発
[分類結果]	: 言語処理 (辞書(25), 自然語(23), LR解析(6)・・・)
[新キーワード]	: 構文解析, 形態素, 統語

【図14】

図14

【画面1】

		正解	不正解
--	--	----	-----

次に示す通りに分類しました。
分類結果が正しければ【正解】、正しくなければ【不正解】をマウスで指定して下さい。

【文書No】 : 101
【題名】 : 機械翻訳用辞書構築ツールの開発
【分類結果】 : 言語処理 (辞書(25), 自然語(23), LR解析(6)・・・)
【分類候補】 : 通信 (ネットワーク(3))

【画面2】

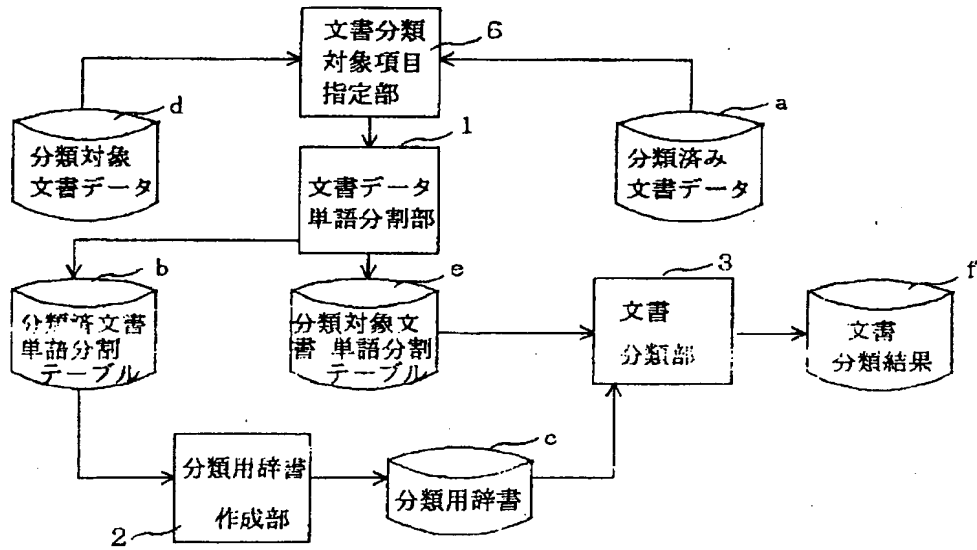
		なし ←
--	--	------

【言語処理】のキーワードとして、誤ったキーワードがあれば該当するキーワードをマウスで指定してください。
(指定されたキーワードは分類用辞書から削除します。)

【キーワード】 : 辞書, 自然語, LR解析, ...

【図16】

図16



【公報種別】特許法第 17 条の 2 の規定による補正の掲載

【部門区分】第 6 部門第 3 区分

【発行日】平成 13 年 2 月 9 日 (2001. 2. 9)

【公開番号】特開平 6-348755

【公開日】平成 6 年 12 月 22 日 (1994. 12. 22)

【年通号数】公開特許公報 6-3488

【出願番号】特願平 5-135588

【国際特許分類第 7 版】

G06F 15/40 500

【F I】

G06F 15/40 500

【手続補正書】

【提出日】平成 12 年 5 月 18 日 (2000. 5. 18)

【手続補正 1】

【補正対象書類名】明細書

【補正対象項目名】特許請求の範囲

【補正方法】変更

【補正内容】

【特許請求の範囲】

【請求項 1】文書データを分類する方法において、
1 つの分類が少なくとも 1 つの文書データからなる分類済みの文書データから分類別のキーワードとなる語抽出して分類用辞書を作成し、
前記分類用辞書を用いて未分類の文書データを分類することを特徴とする文書分類方法。

【請求項 2】文書データを分類する処理システムにおいて、

1 つの分類が少なくとも 1 つの文書データからなる分類

済みの文書データから分類別のキーワードとなる語抽出して分類用辞書を作成する手段、

前記分類用辞書を用いて未分類の文書データを分類する手段を有することを特徴とする文書分類システム。

【請求項 3】前記分類用辞書を作成する手段は、
分類済みの文書データを用いて前記文書データ内の単語を検出して唯一の分類に出現する単語を検出し、前記単語を前記分類を表わすキーワードとして前記分類用辞書に登録することを特徴とする請求項 2 記載の文書分類システム。

【請求項 4】前記未分類の文書データを分類する手段は、
前記未分類文書データ中の単語を検出して前記分類用辞書に登録済みのキーワードとの一致数を検出し、一致した分類の中で最も一致数が多い分類を前記未分類文書データの分類結果とすることを特徴とする請求項 2 記載の文書分類システム。